(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(54) Title: LOAD BALANCING VIA MESSAGE SOURCE SELECTION
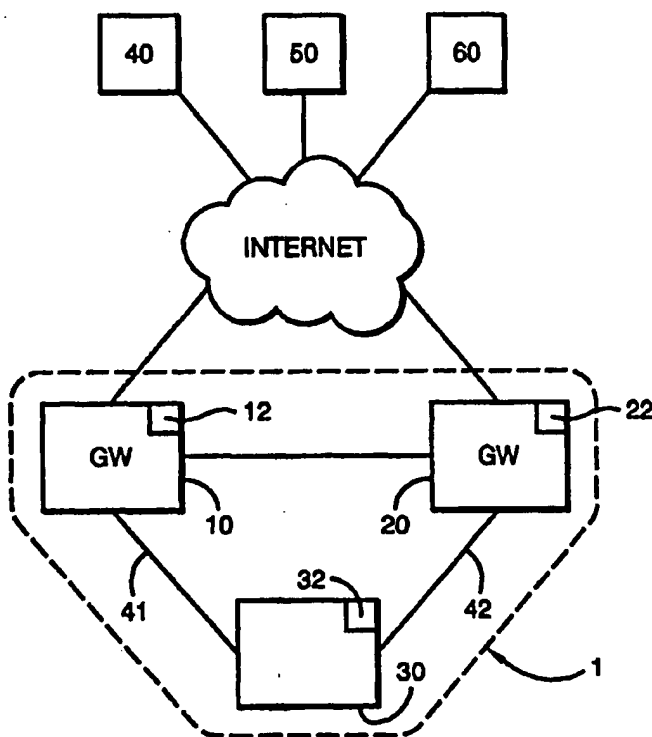
(57) Abstract: Methods and devices for transferring large e-mail messages are described herein. More specifically, one such method comprises providing a mail network comprising an initiator (30) and at least two gateways (10 and 20), the initiator identifying a mail message recipient and identifying advertising material to be mailed to the recipient (40), selecting a gateway from the at least two gateways, and requesting the selected gateway to generate an e-mail message to the recipient; and the selected gateway generating and sending the requested message.

WO 01/27772 A1

(AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Published:
— *With international search report.*
— *With amended claims.*

LOAD BALANCING VIA MESSAGE SOURCE SELECTION

Field of The Invention

5   The field of the invention is computer networks.

Background of The Invention

In a typical electronic mail system, when a user sends a mail message to an external address, the following typically occurs: first, the user interacts with a user agent (such as "sendmail" on UNIX systems or Microsoft Outlook® on Microsoft Windows® systems)

10   which accepts text and a request to send the text; second, upon receipt of the request to send, the text is converted into an electronic mail (e-mail) message and transferred to a local mail transfer agent (MTA) (such as "mailhost" on a UNIX systems or "Microsoft Exchange Server®" on Microsoft Windows® systems); finally, the local MTA, either directly or through a relay MTA, transfers the message to a receiving MTA ("mailhost" or "Microsoft

15   Exchange Server®"). Details regarding mail systems as implemented on TCP/IP networks can be found in the series *TCP/IP Illustrated* published by Addison Wesley Longman, Inc., herein incorporated by reference in its entirety, and particularly Volume 1 of the series by W. Richard Stevens.

The step of transferring the message from the local MTA to the receiving MTA often

20   involves choosing a route from the local MTA to the receiving MTA. Route selection is typically accomplished via low level protocols and "quasi-static" routing tables. Routing tables are "quasi-static" in the sense that they only change when the topography of the network changes, or when manually modified by an operator. Such tables are typically used to implement "cost based" routing in that each link in a table has an assigned cost and route

25   selection is done so as to minimize the cost.

For a system which generates large amounts of long e-mail messages, the local MTA (or the relay MTA if one is used) can become overwhelmed due to the volume of traffic to be handled. For very high volumes of traffic, there may be a significant delay between the time when a message is originally "sent" (given to the user agent) and when it actually gets

30   delivered (transferred by the local or relay MTA to a receiving MTA).

One method that might be applied to help prevent overwhelming an MTA is load balancing. Although load balancing mechanisms are known, they are generally limited to handling high volumes of incoming requests for data files such as requests for "web" pages. In such systems, an incoming request is initially directed to a "traffic cop" such as a load

5   balancing router which then directs the request for data file and/or connection to an available computer or other device. Although such systems may be advantageous in the proper circumstances, there existence does not provide a mechanism for balancing the load caused by a high volume of outgoing mail messages.

A modified form of cost based routing replaces the cost figure with a desired relative

10   frequency as is described in U.S.Patent No. 5,872,930 herein incorporated by reference in its entirety. Although this form achieves a type of load balancing, it is still based on quasi-static tables.

Quasi-static tables are not always sufficient because they do not take into account variations in network traffic or the capabilities of individual links. For example, a route over

15   which a message is to travel based on the routing tables may be overburdened and acting as a bottleneck. Such bottlenecks can be particularly troubling when a mail message includes large amounts of data.

Summary of the Invention

Methods and devices for load balancing via gateway and source selection are

20   described herein. More specifically, one such method comprises: providing a mail network having an initiator and at least two gateways, the initiator identifying a mail message recipient and identifying advertising material to be mailed to the recipient; selecting a gateway from the at least two gateways; requesting the selected gateway to generate an e-mail message which includes the identified advertising material; and the selected gateway generating and sending

25   the requested message to the mail message recipient.

Various objects, features, aspects and advantages of the present invention will become more apparent from the following detailed description of preferred embodiments of the invention, along with the accompanying drawings in which like numerals represent like components.

30

2

Brief Description of The Drawings

Fig. 1 is a schematic view of a network embodying the invention.

Fig. 2 is a diagram of a procedure embodying the invention.

Detailed Description

5   Referring first to figure 1, a load balanced mail network 1 (LBM net) comprises at least two gateways 10 and 20, each gateway having a message transfer agent (12 and 22) capable of generating and sending electronic mail message to a plurality of external computers (40, 50, 60); at least one initiator 30 comprising an initiating agent 32, the initiating agent 32 capable of obtaining information on available gateways (10 and 20),

10   identifying a mail recipient, identifying advertising material to be mailed to the recipient, choosing a gateway (10 or 20) based on the obtained information and/or the identified recipient and advertising material, and transferring a request to generate an electronic mail message to the chosen gateway (10 or 20). The selected gateway generates and sends an e-mail message containing the identified advertising material to the identified recipient's

15   account on one of the external computers (40, 50, or 60).

As used herein, the term "mail network" means the initiator, the gateways, and the communication paths among them. A load balanced mail network is a mail network in which at least some of the various components cooperate to balance the load on the network to maximize throughput. Any computer, routers, bridges, or other devices which are not part of

20   the mail network are referred to as being "external" to the mail network. Thus, an "external computer" is a computer which is not an initiator, a gateway, or part of the communication pathway between the initiator and gateway. External devices may be interconnected so as to form one or more external networks such as the INTERNET. As an example, on an office LAN which has a single computer to connect the LAN to the INTERNET and to host an

25   MTA for accepting e-mail messages from user agents on individual machines, the LAN is the mail network and the INTERNET is an external network comprising many external devices.

The gateways provide a clean point of separation between the mail network and the external network. In most embodiments, the separation will be a physical one in which the only physical connection between the mail network and the external network is the gateway.

30   Thus, in the office LAN example previously discussed, the gateway and any device or

3

communication path on the LAN side of the gateway is part of the mail network and any device or communication path which is on the INTERNET side of the gateway is an external device. However, not all embodiments need have physically separation between the mail network and external network, such as where one or more communication paths in the mail

5   network are shared with the external network. In such embodiments, the separation is based more on separation of device functions and flow of data than on physical separation with any data/packets destined for the external network flowing through the gateway as though the mail network and external network were physically distinct.

As was the case in the office LAN example, it is preferred that the gateway be a

10 computer coupling the mail network to the INTERNET and via the INTERNET to a plurality of recipient mail hosts/destinations. In such an embodiment, the INTERNET is an external network and the recipient mail hosts are external devices.

Gateways 10 and 20 are preferred to be computers which can accept requests from the initiator to assemble and disseminate mail messages from the initiating agent 32 running on

15 the initiator 30, and to perform the requested assembly and dissemination. Gateways 10 and 20 may include any combination of hardware and software which allow them to accomplish their tasks, however, in preferred embodiments, the initiating agent 32 and MTAs 12 and 22 will have a client-server type relationship with the MTAs 12 and 22 acting as the server side.

It is contemplated that the gateways 10 and 20 might each have a set of multimedia

20 files stored on them for inclusion in outgoing mail messages where the set on each gateway is identical with the sets on the other gateways. By storing a set of multimedia files on each gateway, the initiating agent need not send large amounts of data to each gateway, but instead need only send a request specifying the recipient of the message and what to include in the message.

25 One advantage of having the initiating agent send requests to "assemble and mail" rather than complete messages is that it helps prevent the network connecting the initiator to the gateways from carrying excess traffic, as the message is assembled from data located on the gateway. Thus, even though the system as a whole might require gigabytes per second of bandwidth, internal links do not correspond in size as the requests to "assemble and mail"

30 take relatively little bandwidth.

Each gateway is also preferred to be connected to different points of an external network such as the INTERNET. In such a case, any outgoing mail or mail requests could be directed to a particular gateway. If information pertinent to message delivery were maintained for each gateway, it would be possible to use that information for choosing a

5   "preferred" gateway for a given message or message request. Thus, it might be possible to choose a gateway because using that particular gateway will minimize the time needed to transfer a particular message to its recipient. The information collected may include but is not necessarily limited to the following: delivery time, round trip time, gateway response time, and files available on the gateway.

10   The step of having the initiating agent choose the gateway to use from among several possible gateways, can be viewed as providing a second layer of routing to the network. Such application layer routing allows the initial choice of routes to be based on information generally not available to lower level protocols. The ability to perform application layer routing is at least partially supported by maintaining a plurality of gateways having similar

15   capabilities. It is contemplated that the concept of application can have many embodiments. The concept might best be described as, in a network comprising at least two layers or levels of protocols where one layer is a routing layer, using at least one layer other than the routing layer to route packets from one machine to another.

Initiator 10 and initiating agent 12 are preferred to maintain a list of recipients, data

20   indicating preferences of the recipients, and available advertising materials. The initiator periodically sends requests to one of at least two gateways capable of generating e-mail messages containing selected advertising materials and sending such generated messages in response to requests received from the initiator. As with gateways 10 and 20, initiator 30 may comprise a combination of hardware and software which provides the necessary capabilities.

25   As used herein, "advertising materials" includes any material designed to cause someone to think or act in a particular way such as to vote for a particularly political candidate or purchase a particular product. The term includes pictures, graphics, sound, and video either individually or in combination, particularly electronically stored versions of such works. The term "multimedia file" as used herein includes computer files containing one or

30   more such works.

Related methods of routing, load balancing, and additional information on possible relationships between the initiator and gateways can be found in U.S. Patent Applications titled "Message content based routing", "Dynamic Routing via Shortest Delivery Time", and "Historical Delivery Time Based Routing Tables" having inventors identical to this

5   application and filed concurrently with this application, all of which are hereby incorporated by reference in their entireties.

Thus, specific embodiments and applications of methods and devices for load balancing via gateway selection have been disclosed. It should be apparent, however, to those skilled in the art that many more modifications besides those already described are possible

10   without departing from the inventive concepts herein. The inventive subject matter, therefore, is not to be restricted except in the spirit of the appended claims. Moreover, in interpreting both the specification and the claims, all terms should be interpreted in the broadest possible manner consistent with the context. In particular, the terms "comprises" and "comprising" should be interpreted as referring to elements, components, or steps in a non-exclusive

15   manner, indicating that the referenced elements, components, or steps may be present, utilized, or combined with other elements, components, or steps that are not expressly referenced.

CLAIMS

What is claimed is:

1. A method for load balancing via e-mail source selection comprising:

    providing a mail network comprising an initiator and at least two gateways;

    the initiator

    identifying a mail message recipient and identifying advertising material to be
    mailed to the recipient,

    selecting a gateway from the at least two gateways, and

    requesting the selected gateway to generate an e-mail message which includes the
    identified advertising material and to send the generated message to the
    recipient;

    the selected gateway generating and sending the requested message.

2. The method of claim 1 wherein each gateway comprises a set of multimedia files for
    inclusion in generated mail messages, the set on any one gateway including the same
    files as the sets on all the other gateways.

3. The method of claim 2 wherein the initiator comprises a list of the multi-media files
    included in the sets on the gateways, and a database of message recipients includes
    data on each recipient including the identified recipient, the data for the identified
    recipient being used in identifying the advertising material to be mailed to the
    identified recipient.

4. The method of claim 1 wherein the mail network is a WAN.

5. The method of claim 1 wherein the gateways are coupled to the INTERNET such that
    the generated message is sent to the recipient via the INTERNET.

6. The method of claim 1 wherein the selection of the gateway is at least partially based
    on attempting to minimize the time between when the gateway is requested to
    generate the message and when the recipient receives the message.

7.   The method of claim 6 wherein each gateway can determine the round trip time between itself and the recipient's machine and the selection of a gateway is at least partially based on the determined RTT.

8.   A method for load balancing via e-mail source selection comprising:

providing a mail network comprising an initiator and at least two gateways, each gateway comprising a set of multimedia files for inclusion in generated mail messages, the set on any one gateway including the same files as the sets on all the other gateways;

the initiator

comprising a list of the multi-media files included in the sets on the gateways;

comprising a database of message recipients which includes data on each recipient including the identified recipient;

identifying a mail message recipient and using the data for the identified recipient in identifying advertising material to be mailed to the recipient,

selecting a gateway from the at least two gateways wherein the selection of the gateway is at least partially based on attempting to minimize the time between when the gateway is requested to generate the message and when the recipient receives the message, and

requesting the selected gateway to generate an e-mail message which includes the identified advertising material and to send the generated message to the recipient;

the selected gateway generating and sending the requested message.

AMENDED CLAIMS
[received by the International Bureau on 9 March 2000 (09.03.00);
original claims 1 and 8 amended; remaining claims unchanged (2 pages)]

CLAIMS

1. In a client-server architecture. a method for load balancing outgoing e-mail via e-mail source selection comprising:
   providing a mail network comprising an initiator and at least two alternative gateways distinct from the initiator:
   the initiator identifying a mail message recipient and identifying advertising material to be mailed to the recipient. automatically selecting a gateway from the at least two gateways, and requesting the selected gateway to automatically generate an e-mail message which includes the identified advertising material and to send the generated message to the recipient: and
   the selected gateway generating and sending the requested message.

2. The method of claim 1 wherein each gateway comprises a set of multimedia files for inclusion in generated mail messages. the set on any one gateway including the same files as the sets on all the other gateways.

3. The method of claim 2 wherein the initiator comprises a list of the multi-media files included in the sets on the gateways. and a database of message recipients includes data on each recipient including the identified recipient. the data for the identified recipient being used in identifying the advertising material to be mailed to the identified recipient.

4. The method of claim 1 wherein the mail network is a WAN.

5. The method of claim 1 wherein the gateways are coupled to the INTERNET such that the generated message is sent to the recipient via the INTERNET.

6. The method of claim 1 wherein the selection of the gateway is at least partially based on attempting to minimize the time between when the gateway is requested to generate the message and when the recipient receives the message.

7. The method of claim 6 wherein each gateway can determine the round trip time between itself and the recipient's machine and the selection of a gateway is at least partially based on the determined RTT.

8. In a client-server architecture. a method for load balancing outgoing e-mail via e-mail source selection comprising:

9

AMENDED SHEET (ARTICLE 19)

providing a mail network comprising an initiator and at least two alternative

    gateways distinct from the initiator. each gateway comprising a set of

    multimedia files for inclusion in generated mail messages. the set on any one

    gateway including the same files as the sets on all the other gateways:

the initiator comprising a list of the multi-media files included in the sets on the

    gateways: comprising a database of message recipients which includes data on

    each recipient including the identified recipient: identifying a mail message

    recipient and using the data for the identified recipient in identifying

    advertising material to be mailed to the recipient:

selecting a gateway from the at least two gateways wherein the selection of the

    gateway is at least partially based on attempting to minimize the time

    between when the gateway is requested to generate the message and when

    the recipient receives the message. and

requesting the selected gateway to generate an e-mail message which includes the

    identified advertising material and to send the generated message to the

    recipient:

the selected gateway generating and sending the requested message.

10

AMENDED SHEET (ARTICLE 19)

1/1

FIG. 1

FIG. 2

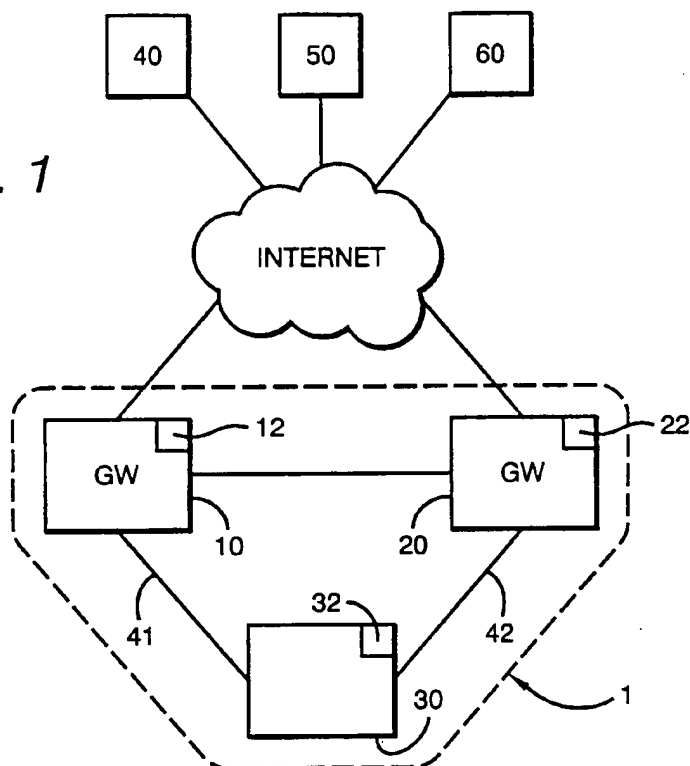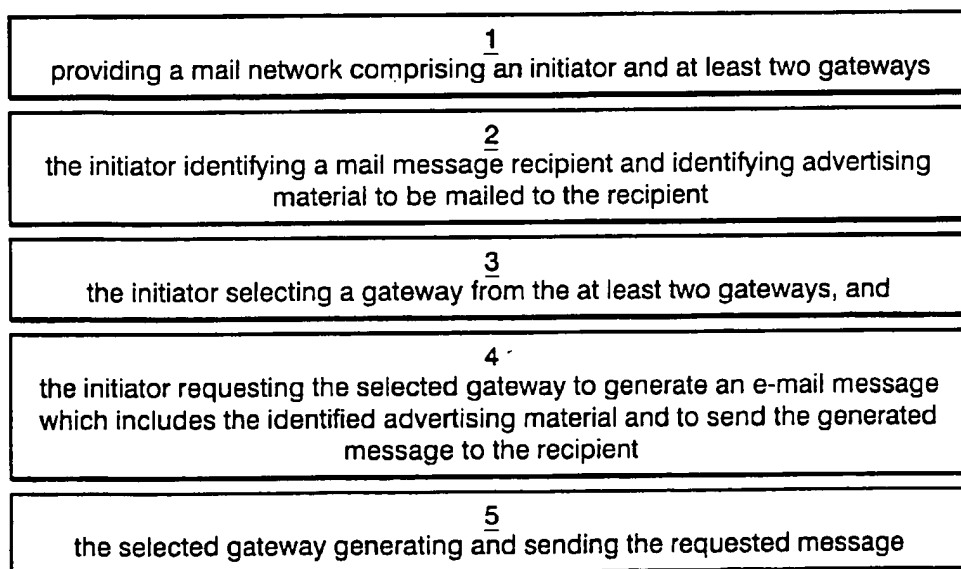| **1**
providing a mail network comprising an initiator and at least two gateways |
|---|
| **2**
the initiator identifying a mail message recipient and identifying advertising
material to be mailed to the recipient |
| **3**
the initiator selecting a gateway from the at least two gateways, and |
| **4**
the initiator requesting the selected gateway to generate an e-mail message
which includes the identified advertising material and to send the generated
message to the recipient |
| **5**
the selected gateway generating and sending the requested message |

SUBSTITUTE SHEET (RULE 26)

INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :G06F 13/00, 13/38
US CL : 709/206, 202, 200, 207; 379/34

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 709/206, 202, 200, 207; 379/34

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

IEEE, MICROSOFT'S COMPUTER DICTIONARY

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

DIALOG, WEST, EAST
search terms: e-mail, lan, internet, gateway, load balance, initiate, routing, advertising, identify

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5,790,789 A (SUAREZ) 04 August 1998, figs. 1 and 5, abstract, col.6 lines 13-41; col.9, line 40 to col.10, line 58; col.11 lines 15-43.	1-8
Y	US 5,632,018 A (OTORRI) 20 May 1997, abstract, fig.2, col.3, lines 3-38; col.4, lines 5-59.	1-6, 8
Y	US 5,765,033 A (MILOSLAVSKY) 09 June 1998, col.2, line 57 to col.3 line 53; col.4, lines 20-59.	7

☐ Further documents are listed in the continuation of Box C.   ☐ See patent family annex.

*	Special categories of cited documents:	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	
"E"	earlier document published on or after the international filing date	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means	
"P"	document published prior to the international filing date but later than the priority date claimed	"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 DECEMBER 1999	19 JAN 2000
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231	Authorized officer AHMAD MATAR
Facsimile No. (703) 305-3230	Telephone No. (703) 305-4731

Form PCT/ISA/210 (second sheet)(July 1992)*